基于微博的细粒度情感分析

敦欣卉[1] 张云秋[1] 杨铠西[2]
[1](吉林大学公共卫生学院 长春 130021)
[2](大连理工大学中日国际信息与软件学院 大连 116620)

摘要: 【目的】对微博进行细粒度情感分析, 将情感分为 8 类, 并计算其情感强度值, 从而尽可能还原微博用户情感。【方法】通过微博语料分析构建疑问词词表, 在大连理工大学情感词汇本体 DUTIR 的 7 类情感基础上, 丰富一类情感"疑", 并利用点互信息法构建表情符号词典, 还综合考虑否定词和程度副词对情感表达的影响, 利用 Python 从新浪微博上获取数据, 并用 R 语言的 jiebaR 包进行分词, 对情感进行分类并计算其强度。【结果】得到微博用户对于糖尿病 7 类常用药物的 8 类情感占比及情感强度, 并通过正确率、召回率、F 值对结果进行验证, 其中"怒"和"哀"的正确率最高, 分别为 85.73%和 83.05%, 而"乐"和"好"的召回率与 F 值均最高, 为 81%以上。本文新增情感"疑"的正确率、召回率、F 值分别为 77.33%、78.58%、77.95%, 均值在 8 类情感中排名前列, 说明其情感识别较好。【局限】由于本文依赖于情感词典进行情感分析, 因此为了更好的分析结果, 情感词典仍需进一步完善。【结论】本方法具有较高的识别率和可靠性, 能够更好地对微博上的情感分类进行细粒度分析。

关键词: 微博 细粒度情感分析 药物
分类号: TP393

1 引 言

中国互联网络信息中心于 2017 年 1 月发布的《中国互联网络发展状况统计报告》显示, 截至 2016 年 12 月, 中国网民规模达 7.31 亿, 互联网普及率为 53.2% [1], 人们对网络的利用率越来越高。随着 Web 3.0 技术的发展, 互联网上出现了社区、论坛、博客、微博等各种形式的社会化媒体平台, 它们帮助用户在网上表达自己对某一事件的看法, 使人们通过互联网相互影响。其中微博(Microblog)具有用户多、消息数量大、更新快等特性, 成为人们获取信息、发表舆论的主要途径, 越来越多的明星、政府机构、企业等也选择微博进行重要信息的发布和传播, 这些信息充斥了大量的社会热点及情感。通过对微博用户发布的内容进行细粒度情感分析, 尽可能还原用户真实情感, 有助于人们及时获取热门话题, 帮助控制社会舆论走向, 也有助于对产品评论进行分析, 不仅能够辅助用户优化自身的购买决策, 还能够帮助企业有针对性地进行自我改进, 提升市场竞争力, 准确地发现并挖掘微博中潜藏的商业价值和社会价值。

2 相关研究

微博情感分析是指通过分析和挖掘微博中的主观性信息来判断其情感倾向。目前国内已有较多关于微博情感分析的研究, 按其粒度可划分为两大类, 粗粒度的情感分析和细粒度的情感分析。粗粒度的情感分析主要是基于篇章级和句子级, 而且在分析过程中仅考虑情感词, 并未考虑评价对象及其属性的情感; 细粒度的情感分析一般指词汇级情感分析, 目前关于细粒度情感分析的研究主要分为两大方面: 一方面是文本中产品属性和对应情感词的抽取, 另一方面是对情感进行分类。在产品属性的提取方面, 主要有三种方法, 一种是基于人工定义的方法, 需要针对特定领域的产品建立该领域的产品属性词汇表或产品本体[2], 如李长江构建了一个酒店领域的特征词典, 并在常用的中文情感词典的基础上抽取酒店领域评论中的情感

通讯作者: 张云秋, ORCID: 0000-0002-9790-9581, E-mail: yunqiu@jlu.edu.cn。

词构建情感词典[3]；另外一种是基于自动提取的方法，通过词性标注、句法分析等自然语言处理技术对产品评论中的语句进行分析，从中自动化提取产品属性[2]，如贾治中在依存句法分析的基础上添加一系列语义规则，显著提高了评价对象的抽取性能[4]；还有一种是使用主题模型的方法，如彭云等提出语义关系约束的主题模型 SRC-LDA，用来实现语义指导下 LDA 的细粒度主题词提取[5]。在情感分类方面，无论是粗粒度还是细粒度的情感分析，所用的方法均可分为三类，有监督的机器学习方法、无监督情感分析方法和半监督情感分析方法。有监督机器学习方法通过选取例如情感词等的情感分类特征，通过分类器完成有监督的训练和测试。具有里程碑意义的是 Pang 等应用三个代表性分类器(支持向量机 SVM、朴素贝叶斯 NB、最大熵 ME)对文本进行情感分类，得出机器学习的文本情感分类性能较好，可达到 80%的准确率[6]；还有学者对不同的分类算法进行比较，杨艳霞利用贝叶斯算法和 SVM 分类算法对微博进行情感分析，并比较了两种算法在分类性能上的优劣，从而得出贝叶斯算法的准确性更高[7]；还有学者对分类算法进行改进，从而使分类效果更好，陈炳丰等对 Linear-chain CRF 模型进行改进，提出一种双层结构的 CRF 模型，从而能够更好地满足汽车评论在情感实体识别与情感倾向分类的需求[8]；半监督分析方法基于小部分已标注数据集，通过对部分无标注数据进行测试来扩大已标注数据集规模，之后进行迭代，逐步预测数据。朱晓光[9]结合已有的标注集运用半监督学习中的主动学习方法标注微博文本的情感极性和类别，以减少标注成本，并将标注的数据集应用于监督学习中；程佳军[10]提出基于半监督递归自动编码的微博文本情感分类方法，对微博进行情感分析，并在多个数据集上较基于支持向量机的文本情感分类方法取得了更好的效果。但由于半监督分析方法初始标注规模小，其最终学习性能也持续削弱，因此不具备高精度能力。

由于有监督学习依赖于充足的标注语料，但是微博这种数量庞大的互联网文本导致人工不能标注大规模的语料，其适用领域与规模受到限制。此外，由于微博中蕴含了表达情感倾向的多种表情符号和网络用语，对其进行标注时也容易受到符号变形、种类的制约，因此，基于有监督方法的情感分类并不适用于微博，微博中情感分类的研究更多倾向于没有标注样本的无监督学习方法。

无监督情感分析方法主要基于现有的情感词典或者对已有的情感词典扩充来对文本进行情感分析。目前有代表性且使用较广泛的词典资源，英文领域主要有 WordNet、General Inquirer 等。中文领域常用的情感词典有《知网》(HowNet)、NTUSD、C-LIWC、DUTIR 等。熊德兰等基于 HowNet 对句子的褒贬性进行了研究[11]；潘明慧等提出了基于词典的方法识别出微博表达的 6 种情绪[12]。情感词典扩充的方法主要分为两部分：一部分利用特定领域语料构建适用于该领域的词典，如肖江等利用基于知网的语义相似度算法在 HowNet 的基础上构建领域情感词典，使基础情感词典不适用于领域情感分析的问题得到一定的改善[13]；另一部分通过计算未登录词与已知情感类别词的语义相似度来进行扩充，如王志涛等基于新浪微博平台利用统计信息和点互信息法识别新词及情感标注，最终构建了微博新词情感词典[14]。近年来，随着微博情感分析研究的深入，越来越多人将目光转向其他表达情感的情感元素的词典构建上，例如张珊等利用微博中的表情图片并结合情感词语的方法构建了中文微博情感语料库[15]；王文远等构建了一种表情符号词典将文本分为正负性[16]；栗雨晴等构建了中英文双语词典将文本分为 5 类情感，结果表明其准确率高于传统的分类方法[17]。

虽然目前已有众多从方法及应用[18-22]的角度对微博进行细粒度情感分析的研究，学者们在对文本中评价对象及其特征和对应的情感词进行提取方面取得了一定的进步，但对于情感的分类多是基于正负二元、或者加上中性三元分类，对于情感的分类较粗且没有考虑情感强度。人类情感复杂，对其情感的分析研究不能只停留在好恶层面，应尽可能细分情感类别并且计算情感强度，从而在真实还原人类情感的基础上进行相关研究。虽然缪茹一[23]、崔安颀[24]等少数学者也将情感进行了喜怒哀乐等细致分类，但是均不涉及“疑”这类情感。对人类情感的缺失识别并不能满足人们对于情感分析的需求。此外，人们对于细粒度情感分析的研究多局限于情感分类，并没有计算其情感强度值。而情感必然会伴随着强弱的表达，缺失了情感强度值的比较，情感分析也并不完善。因此，本文在对

于微博的情感进行情感分析时，不仅通过情感词词典进行情感分类，还考虑到同样具有情感表达作用的表情符号，利用点互信息法构建了表情符号词典，在大连理工情感词汇本体库 DUTIR 的"乐、好、怒、哀、惧、恶、惊"7 类情感基础上增加了"疑"类情感，并考虑到程度副词与否定词对于情感表达的影响，将其作为影响因素对每类情感的情感强度进行计算，从而更细腻地分析微博中的情感，有助于人们的后续研究。

3　情感分析流程与方法

3.1　情感分析流程

本文在大连理工大学情感词汇本体库的"乐、好、怒、哀、惧、恶、惊"7 类情感基础上丰富了一类表示疑惑的"疑"情感，将用户的情感分为 8 类，并且为了更准确地计算出每类情感的情感强度，还利用点互信息法(PMI)构建了表情符号词典，此外，还综合考虑了否定词及程度副词等修饰词对于情感词的影响，构建了程度副词词表和否定词词表，并将其赋予一定权重，以便于情感强度的计算。以微博上 2 型糖尿病 7 类常用药物数据为例，利用Python从新浪微博上获取数据，并用 R 语言中的 jiebaR 包进行分词，结合所构建的词典，得到微博用户对于药物的细粒度情感分析，并利用正确率、召回率以及 F 值对结果进行验证。此外，为了更好地对药物进行比较，利用 R 语言对切词后的数据进行统计，得到能够代表用户所关心的药物的高频特征，并对其进行情感分析，从而得知用户对于药物某类特征的情感倾向及强度。微博情感分析流程如图 1 所示。

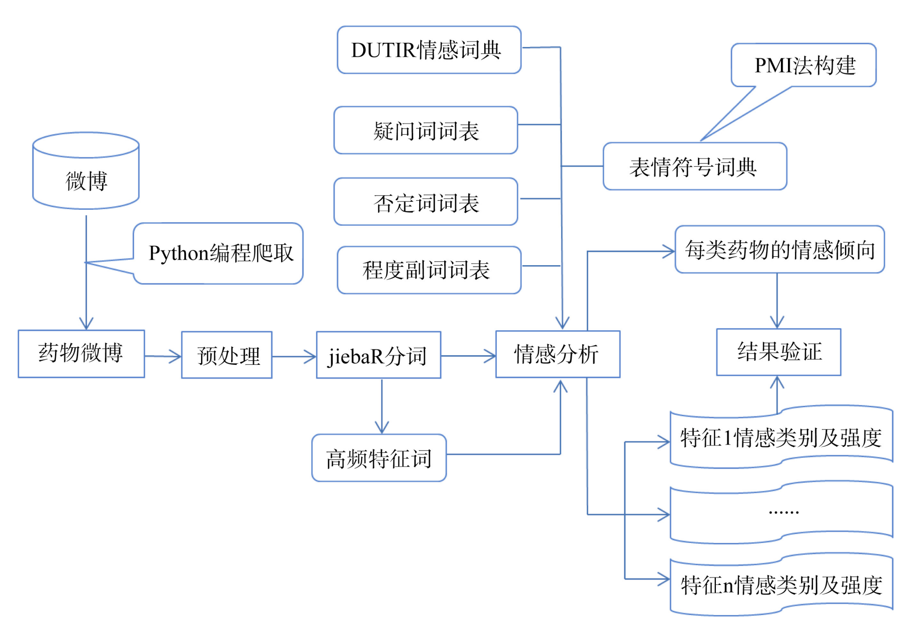


图 1　微博情感分析流程

3.2　研究方法

(1) 数据的获取与处理

利用 Python 语言进行编程，以完成整个数据的获取，获取字段包括微博文本(text)、评论数(comment)、转发数(transfer)、点赞数(like)和用户 ID(uid)。

在大数据环境下，微博由于其社会化媒体的特殊性，其数据鱼龙混杂，会影响情感分析的结果，因此，需要对获取的微博数据进行一些必要的处理。数据清洗规则如下:

①删除与目标内容无关的微博;

②删除因转发而重复爬取的微博，只留取其中一条;

③改正微博中繁体字、错别字等。

由于情感分析依赖于情感词典，因此必须对清洗后的数据进行分词。由于 R 语言的分词包 jiebaR 词汇量大且一直处于更新状态中，其分词准确，处理速度快，并且能够支持用户词典，因此本文采用 jiebaR 作为中文分词工具。

(2) 基于 DUTIR 的情感补充

DUTIR(中文情感词汇本体库)是大连理工大学信息检索研究室整理和标注的一个中文本体资源[25]。词

汇本体中的情绪共分为 7 种: “好、乐、哀、怒、惧、恶、惊”, 共含有情绪词 27 466 个, 情感强度分为: 1, 3, 5, 7, 9 这 5 档, 9 表示强度最大, 1 为强度最小。该资源从不同角度描述一个中文词汇或者短语, 包括词语词性种类、情感类别、情感强度及极性等信息。每个词在每一类情感下都对应一个极性。其中, 0 代表中性, 1 代表褒义, 2 代表贬义, 3 代表兼有褒贬两性。为了根据词汇的情感强度值计算微博的情感强度, 本文将褒义极性值不变, 贬义极性值取–1, 如表 1 所示。

表 1　情感词汇本体格式举例

词语	词性种类	词义数	词义序号	情感分类	强度	极性	辅助情感分类	强度	极性
无所畏惧	idiom	1	1	PH	7	1			
手头紧	idiom	1	1	NE	7	0			
周到	adj	1	1	PH	5	1			
言过其实	idiom	1	1	NN	5	–1			

DUTIR 将情感分为 7 大类 21 小类, 如表 2 所示。

表 2　情感分类

编号	情感大类	情感类	例词
1	乐	快乐(PA)	喜悦、欢喜、笑眯眯、欢天喜地
2		安心(PE)	踏实、宽心、定心丸、问心无愧
3	好	尊敬(PD)	恭敬、敬爱、毕恭毕敬、肃然起敬
4		赞扬(PH)	英俊、优秀、通情达理、实事求是
5		相信(PG)	信任、信赖、可靠、毋庸置疑
6		喜爱(PB)	倾慕、宝贝、一见钟情、爱不释手
7		祝愿(PK)	渴望、保佑、福寿绵长、万寿无疆
8	怒	愤怒(NA)	气愤、恼火、大发雷霆、七窍生烟
9	哀	悲伤(NB)	忧伤、悲苦、心如刀割、悲痛欲绝
10		失望(NJ)	憾事、绝望、灰心丧气、心灰意冷
11		疚(NH)	内疚、忏悔、过意不去、问心有愧
12		思(PF)	思念、相思、牵肠挂肚、朝思暮想
13	惧	慌(NI)	慌张、心慌、不知所措、手忙脚乱
14		恐惧(NC)	胆怯、害怕、担惊受怕、胆颤心惊
15		羞(NG)	害羞、害臊、面红耳赤、无地自容
16	恶	烦闷(NE)	憋闷、烦躁、心烦意乱、自寻烦恼
17		憎恶(ND)	反感、可耻、恨之入骨、深恶痛绝
18		贬责(NN)	呆板、虚荣、杂乱无章、心狠手辣
19		妒忌(NK)	眼红、吃醋、醋坛子、嫉贤妒能
20		怀疑(NL)	多心、生疑、将信将疑、疑神疑鬼
21	惊	惊奇(PC)	奇怪、奇迹、大吃一惊、瞠目结舌

人类是不断探索的生物, 无论是对于他人的咨询, 还是对未知世界的探索, 表达疑问、困惑的“疑”类情感在人类全部情感中占有相当的比例。尤其当今是网络时代, 人们通过社会化媒体进行信息的搜寻或者浏览时, 不仅传统的表达疑问的“为什么”等疑问词比比皆是, 表达疑问的“怎么破”等网络用语也随处可见。由于 DUTIR 中没有表示疑问的词汇, 而微博中用户表达疑问的情绪也较多, 因此, 基于《现代汉语词典》与新浪微博, 笔者搜集构建了一个常见疑问词词表, 共 52 个疑问词, 如表 3 所示。将其按照表达强弱, 分为 4 个等级, 其极性与情感强度值依据 DUTIR 格式由人工标注, 作为 DUTIR 的补充情感。其中, 由于疑问词词典是为了分析微博中的疑问情绪, 因此, 所有疑问词极性值均取 1, 便于后续计算。

表 3　疑问词词表

序号	疑问词	强度值	极性值
1	哪儿、哪里、怎么样、怎么着、如何、为什么、难道、'呢？'、'吧？'、'啊？'、啥、为何、怎么办、哪些、问题、请问、为神马、神马情况、为啥、干嘛、能否、何时、求问	7	1
2	谁、何、什么、神马、几时、怎么、怎的、怎样、岂、何尝、吗、么、多大、有没有、会不会、好不好、能不能、可不可以、行不行	5	1
3	几、多少、怎、难怪、反倒、何必、你知道	3	1
4	居然、竟然、究竟	1	1

(3) 修饰词词典的构建

用户对于情绪的表达往往不只是含有情感词汇, 还含有大量的副词对情感词汇进行修饰。为了更好地识别微博的情感及其强度, 还需要构建程度副词和否定词等修饰词词表。根据《现代汉语词典》以及前人研究[26], 将程度副词分为 4 个等级: 极量级、高量级、中量级、微量级, 并且参考众多学者对于程度副词权

值的定义方法[27-28]，最终将程度副词的强度取值范围限定在[0, 2][29]，按照 4 个等级递减强度值，强度值越靠近 0，强度越弱，反之则强度越强。最后构建了 51 个程度副词，44 个否定词，如表 4 和表 5 所示。

表 4　程度副词词表

序号	程度副词	强度值
1	极、极为、极其、透顶、极端、顶、最、最为、绝顶、无比	2
2	多、很、非常、甚至、十分、太、分外、特别、万分、尤其、真、格外、何等、过于、多么、更加、更为、更、越加、越发、愈加、愈、相当、好	1.5
3	颇、挺、比较、较、较为、较比	1.2
4	怪、有点、有点儿、有些、稍、稍稍、稍微、稍许、少许、略、略微	0.5

表 5　否定词词表

否定词
白白、甭、别、并非、不、不必、不曾、不可、不要、不用、从不、从未、非、毫不、毫无、何必、何曾、何尝、何须、决不、绝不、绝非、绝无、没、没有、莫、难以、切勿、尚未、徒、徒然、枉、未、未必、未曾、未尝、未有、无从、无须、无庸、毋须、毋庸、勿

(4) 表情符号词典的构建

微博平台上，系统为用户准备了丰富的表情符号以表达他们的情绪，研究显示，含有表情符号的微博占比约为 18.73%[30]，因此表情符号对于微博用户情感展示的作用不容忽视。在爬虫过程中，表情符号会转变为表情符号的 alt 标签所标记的文本内容，如 对应的为[哈哈]， 对应的为[泪]。

虽然新浪微博表情众多，但不是每一个都为人们常用，因此本文选取微博上使用频率最高的 113 个表情符号构建表情符号词典。词典的构建分为两部分：

①将表情符号的 alt 标签内的词与 DUTIR 对应，若找到对应，则将该表情符号划分到该情感词的分类中；

②若未找到对应的表情符号，则利用 PMI 法寻求与之共现频率最高的情感词或已知分类的表情符号，从而将其归为一类。

PMI 法主要用于计算的语义相似度，基本思想是统计两个词语在文本中同时出现的概率，如果概率越大，其相关性就越紧密，关联度越高。两个词语之间，即 word1 和 word2 之间的 PMI 计算公式[31]如下。

$$PMI_{(word1,word2)} = \log_2\left(\frac{P_{(word1\&word2)}}{P_{(word1)}P_{(word2)}}\right)$$

其中，$PMI(word1\&word2)$ 表示两个词语共同出现的频率，$P(word1)$ 和 $P(word2)$ 表示两个词分别出现的频率。若计算值越大，表明两个词语的共现频率越高，相关度越大；反之，则越小。本文将两个词语中的一个词替换为表情符号的 alt 标签值进行计算。

通过这两部分筛选，113 个常用表情符号中，已找到对应的有 74 个，未找到对应的有 39 个。因此利用 Python 语言编程，从新浪微博上爬取含有这 39 个表情符号的数据共为 48 827 条，利用点互信息法，得到表情符号词典如表 6 所示。

表 6　表情符号词典(部分)

表情符号	情感分类	表情符号	情感分类
[doge]	8	[抱抱]	2
[喵喵]	1	[坏笑]	1
[二哈]	1	[舔屏]	2
[打脸]	4	[污]	1
[哆啦 A 梦笑]	1	[允悲]	4
[哆啦 A 梦汗]	7	[笑而不语]	1
[话筒]	2	[费解]	8
[哆啦 A 梦开心]	1	[憧憬]	2
[笑 cry]	1	[并不简单]	2
[摊手]	8	[微笑]	1

最终得到表情符号词典对应情感分类情况如表 7 所示。

表 7　表情符号词典情况

情感分类	表情符号	数量
乐	[微笑][哈哈][偷笑][太开心]	32
好	[爱你][亲亲][鼓掌][心]	31
怒	[怒][抓狂][怒骂]	9
哀	[允悲][委屈][失望][悲伤]	14
惧	[害羞][哆啦 A 梦害怕][羞嗒嗒]	8
恶	[坏笑][挖鼻][闭嘴][鄙视]	8
惊	[吃惊][惊恐]	5
疑	[费解][疑问]	6
总计		113

(5) 微博细粒度情感计算

在微博数据获取过程中，已通过“{…}, ”，对每一

条微博进行分割，因此可将每一条微博看作是一句独立的话，将其进行分词后，依据已经补充完的情感词典及构建好的修饰词词表，就可以快速、精准地运算出微博的情感倾向。本文中每条微博用 Item1、Item2、…、Itemn 表示。由于 DUTIR 中有的情感词汇不只有一个情感分类及强度，因此，对于该情感词 i 的情感强度值，本文用以下公式计算。

$$p_i = \sum_{k=1}^{n} \alpha_k \beta_k \quad (1 \leqslant k \leqslant n, n \in \{1,2\})$$

其中，α 为情感词的情感强度值，β 为情感词的极性值，n 为情感词有几类情感分类，若 n=1，则该情感词只有一类情感，若 n=2，则该情感词有辅助情感分类。

由于 DUTIR 将情感词分为 21 小类，而本文只需最后判别出微博情感的 8 大类，因此需要将情感词的情感分类先归为 8 大类中的某类，方法如下。

$$E_{pi} = \begin{cases} M, |\alpha_{k1}\beta_{k1}| < |\alpha_{k2}\beta_{k2}| \\ N, |\alpha_{k1}\beta_{k1}| \geqslant |\alpha_{k2}\beta_{k2}| \end{cases}$$

其中，M 为情感词汇第一个情感分类，$\alpha_{k1}\beta_{k1}$ 为该类别下的情感强度值，N 为该情感词的辅助分类，$\alpha_{k2}\beta_{k2}$ 为该类别下的情感强度值。M 和 N 均可通过计算机依据表 2 进行映射后得到其具体情感类别。

在得到某个情感词汇的情感分类及情感强度值后，结合所构建的程度副词词表和否定词词表对微博进行情感分析。虽然否定词和程度副词共现很常见，但是二者共现时的位置对于情感表达的影响也应值得注意[26]。情感词、否定词、程度副词的组合模式一般如表 8 所示。

表 8　含情感词的组合模式[32]

序号	类型	示例
1	仅含情感词	热情
2	否定词+情感词	不 热情
3	程度副词+情感词	太 热情
4	否定词+程度副词+情感词	不 太 热情
5	程度副词+否定词+情感词	太 不 热情
6	否定词+否定词+情感词	没有 不 热情

结合程度副词与否定词，情感词的情感值计算公式[32]如下。

$$E_i = (-1)^{O_i} a_i p_i m$$

其中，E_i 为情感词组合的情感值，O_i 代表组合中否定词的数目，a_i 代表组合中程度副词的强度，m 代表组合权值，由于组合 4 的特殊性，设组合 4 的权值为 m=0.4，其他组合权值为 m=1，不起加强或削弱的作用[32]。

由于含有表情符号的微博占比为 18.73%，约为 20%，因此将表情符号词典的权值赋值为 0.2，每条微博中某 j 类情感的表情符号数目为 S_j，考虑到表情符号前也有可能有否定词和程度副词的修饰，则每条微博中表情符号所表达的该类情感的强度值计算公式如下。

$$E_{emoji} = 0.20(-1)^{O_i} a_i m S_j$$

则每条微博中某 j 类情感的强度值计算公式如下，其中 N 为情感词数目。

$$E_j = \sum_{i=1}^{N} E_i + E_{emoji}$$

最终，该条微博的情感分类为 $|E_j|_{\max}$ 所属那类情感，强度值 E= $||E_j|_{\max} - |E_j|_{else}|$，值的符号同 $|E_j|_{\max}$。

由于社会化媒体互动性强的特征，每条微博通常会有评论、点赞和转发，这些行为在一定程度上也代表了该条微博所表达情感的强度，因此本文赋予其权值如下：评论数 x=0.02，点赞数 y=0.1，转发数 z=0.2，综上所述，则每条微博 Item 的情感值计算公式如下。

$$E_{Item} = E \times (1 + 0.02x + 0.1y + 0.2z)$$

4　实证研究

4.1　数据收集与预处理

网络购物已成为当今社会主流的购物方式，而随着社会化媒体技术的进步，人们更倾向于在购物时参考他人对商品的评论以进行决策制定。当今社会医药市场不断发展，治疗同一疾病的药物数量众多、种类繁杂，因此如何进行药物的选购是一个亟待解决的问题。微博已成为人们交流信息的首选平台，人们在这里分享自己或亲人、朋友的用药体验，形成来自用户的用药反应的第一手资料。通过对微博平台上的药物相关微博进行情感分析，不仅有助于为用户选购药品提供可靠的参考，也有助于医药企业及时获取消费者对其产品的评价，以便发现产品的不足之处，采取有效措施提高药品质量，形成一个良好的评价信息系统。现代社会人们的生活模式发生了巨大改变，糖尿病发病率逐年上升。资料显示，2 型糖尿病的发病呈逐

渐加重的流行趋势，并且儿童以及青少年 2 型糖尿病发病人数近年来迅速上升[33]。本文使用 Python 语言进行编程，从新浪微博平台爬取微博用户所发表的 2 型糖尿病的 7 类常用药品相关微博进行分析，分别是双胍类口服降糖药、磺脲类口服降糖药、非磺脲类口服降糖药、α 葡萄糖苷酶抑制剂、胰岛素增敏剂、DPP-4 抑制剂、复方制剂。分别选取每一类药物的药品名称作为关键词，爬取内容包括微博文本(text)及其评论数(comment)、转发数(transfer)、点赞数(like)、用户 ID(uid)，进行清洗后数据为 1 704 条，如表 9 所示。

表 9　药品数据

种类	名称	数量	总计
双胍类口服降糖药	二甲双胍	248	353
	格华止、美迪康	105	
磺脲类口服降糖药	格列吡嗪	119	166
	瑞易宁	47	
非磺脲类口服降糖药	瑞格列奈	162	203
	诺和龙	41	
α 葡萄糖苷酶抑制剂	阿卡波糖	172	260
	拜糖平	88	
胰岛素增敏剂	罗格列酮	61	205
	文迪雅	144	
DPP-4 抑制剂	西格列汀	186	305
	捷诺维	119	
复方制剂	消渴丸	212	212
总计			1 704

4.2　数据结果分析

通过上述情感分析方法对所获取的 1 704 条关于 2 型糖尿病 7 种药物的微博进行分析，结果如图 2 所示。

情感类别	双胍类口服降糖药				磺脲类口服降糖药				非磺脲类口服降糖药				α 葡萄糖苷酶抑制剂				胰岛素增敏剂				DPP-4抑制剂				复方制剂			
	条数	分值	均值	百分比	条数	分值	均值	百分比	条数	分值	均值	百分比	条数	分值	均值	百分比	条数	分值	均值	百分比	条数	分值	均值	百分比	条数	分值	均值	百分比
乐	19	212	11.158	8.12	9	55	6.111	8.3	6	22	3.667	1.63	10	3	0.3	0.2	30	430	14.333	38.15	20	267	13.35	14.34	16	276	17.25	17.57
好	129	1368	10.605	52.38	30	101	3.367	15.21	77	595	7.727	43.98	75	570	7.6	37.19	49	336	6.857	29.82	113	1077	9.531	57.86	65	602	9.262	38.32
怒	1	−14	−14	0.54	1	2.5	2.5	0.38	0	0	0	0	0	0	0	0	0	0	0	0	2	6	3	0.32	3	7	2.333	0.45
哀	7	−18	−2.571	0.69	4	−12.5	−3.125	1.9	0	0	0	0	8	−35	−4.375	2.28	4	8	2	0.71	8	−35	−4.375	1.88	8	−39	−4.875	2.48
惧	5	−16	−3.2	0.61	2	−4	−2	0.6	0	0	0	0	1	−6.5	−6.5	0.42	4	−38	−9.5	3.37	2	−11.5	−5.75	0.62	1	−6	−6	0.38
恶	39	−261	−6.692	9.99	24	−128	−5.333	19.31	30	−176	−5.867	13	48	−218.4	−4.55	14.25	24	−56	−2.333	4.97	22	−102.5	−4.659	5.51	55	−262	−4.764	16.68
惊	0	0	0	0	0	0	0	0	0	0	0	0	4	24.5	6.125	1.6	1	10	10	0.89	1	11	11	0.59	0	0	0	0
疑	64	722.5	11.289	27.67	43	360	8.372	54.3	53	560	10.566	41.39	71	675.4	9.513	44.1	29	249	8.586	22.09	35	351.5	10.042	18.88	36	379	10.528	24.12
总计	264	2611.5	9.892	1	113	663	5.867	1	166	1353	8.151	1	217	1532.8	7.064	1	141	1127	7.993	1	203	1861.5	9.17	1	184	1571	8.538	1

图 2　2 型糖尿病 7 类药物情感分析

(注：①条数=清洗后微博条数–情感值为 0 条数；②均值指的是情感强度均值，计算方式为分/条数；③百分比为各情感分值占总情感值的百分比。)

由于各类药物获取的微博数目不一，仅从情感强度值比较分析情感倾向有失偏颇，因此，图 2 中给出情感强度均值。横向比较来看，微博用户对于 2 型糖尿病 7 类药物的情感，以“怒”最少，对双胍类口服降糖药呈现出“不怒”这个情感；“乐”和“好”的情感比较多且强烈，“哀”和“惧”类虽也占有一定比重，但除了胰岛素增敏剂的“哀”为正值外，其余均为负值或零，说明一部分用户对这些类药物既不喜欢也不讨厌，对于这部分用户，药企可有针对性地进行关注，努力将其转化为积极情绪；值得一提的是，各类药物中“疑”的情感占比并不少，说明人们对 7 类药物均存在一定的疑问；对于“惊”这类情感，7 类药物中呈现两极分化，有的药物为 0，非 0 药物则情感强度值较大。此外，双胍类口服降糖药的情感微博条数最多，可见人们对双胍类口服降糖药情感丰富，且强度较大，说明人们讨论该类药物较其他药物频繁，情感表达丰富。

根据 8 类情感所占的百分比绘制成图，可以清晰地显示人们对 7 类药物的情感倾向，如图 3 所示。

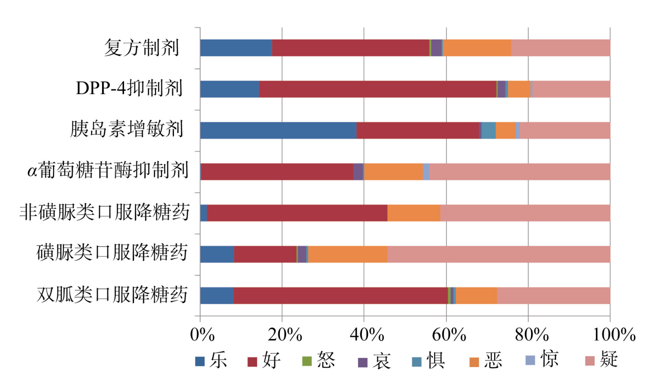


图 3　7 类药物情感倾向分布

由图 3 可知，双胍类口服降糖药、胰岛素增敏剂、DPP-4 抑制剂及复方制剂的“乐”和“好”情感占比较高，说明微博用户比较偏爱于这 4 类药物，其中 DPP-4 抑制剂居首；对复方制剂、非磺脲类口服降糖药、α 葡萄

糖苷酶抑制剂和磺脲类口服降糖药的"恶"的情感占比较高；对复方制剂、α 葡萄糖苷酶抑制剂、DPP-4 抑制剂和磺脲类口服降糖药的"哀"情感占比较高，说明人们对复方制剂、α 葡萄糖苷酶抑制剂和磺脲类口服降糖药这三类药物持消极态度较多；与此同时可以看出人们对 DPP-4 抑制剂的评价呈现两极分化，对复方制剂这类药物情感种类丰富，占比也较多，说明人们对其争议较大；胰岛素增敏剂的"惧"占比最高；而各类药物的"疑"占比均不低，以磺脲类口服降糖药居首，说明人们对于磺脲类口服降糖药的了解不如其他药物，不确定性较多，药企可着重努力改善。

在利用 jiebaR 包对数据进行分词后，对切词结果按照词频排序，并将一些无意义的词过滤掉，最终找出频次大于等于 20 的特征词，共 17 个，如表 10 所示。

表 10 高频特征词表

序号	特征词	词频	序号	特征词	词频
1	糖尿病	145	10	服药	29
2	患者	121	11	第一口	28
3	服用	89	12	餐前	26
4	治疗	84	13	餐后	25
5	降糖药	76	14	用药	25
6	胰岛素	59	15	长生不老	21
7	口服	55	16	副作用	20
8	低血糖	50	17	首例	20
9	餐后血糖	35			

由表 10 可知，人们对于 2 型糖尿病药物多关心其类似于"服用"、"口服"、"餐前"、"餐后"等服用方法以及"副作用"、"低血糖"等药物的副作用。此外，还讨论了药物疗效之外的对人们有利的作用诸如"长生不老"等。为了更进一步了解人们对于这些高频特征词的情感倾向，本文分别将包含每一个高频词的微博数据提取出来，再次进行情感分析。值得一提的是，由于中文的复杂性，微博中与高频词表达同一意思的词仍有许多，因此在采集微博数据时，将特征词的不同表达词汇也扩展进来，使关于特征的情感更为可靠。情感分析结果如图 4 所示。

由图 4 可知，人们谈论药物时的总体情感主要是"好"和"疑"，其次是"乐"，除含有疑问的情绪外，对于药物的总体情感较倾向于乐观；17 个高频特征词中均不存在"怒"，说明在 2 型糖尿病药物中，人们对于药物相关的情感并没有愤怒的情绪在里面；"好"占比最多的特征是"首例"、"长生不老"，说明人们对于 2 型糖尿病药物的首要地位是认可的，对 2 型糖尿病药物能使人延长寿命这一特点多数持积极态度，但也存在少部分的怀疑；对于服药方法中"口服"的情感倾向也好过其他的服药方法。"恶"的占比前三是"降糖药"、"患者"、"餐后血糖"，说明人们仍然对患有 2 型糖尿病的患者需服用降糖药的事实比较反感，对监测餐后血糖这种行为比较厌恶；"惧"的占比最高的是"治疗"和"降糖药"，"哀"最高的是"低血糖"，可知纵然治疗 2 型糖尿病的药物众多，人们还是对降糖药物存在一些恐惧，在药物引起的副作用中对低血糖最为反感；横向比较来看，对于"服用"、"口服"、"餐前"、"餐后"这种表达药物服用方式的词汇，人们的情感"好"的占比有绝对优势，说明人们谈及 2 型糖尿病药物时，更喜欢讨论其服用方法。

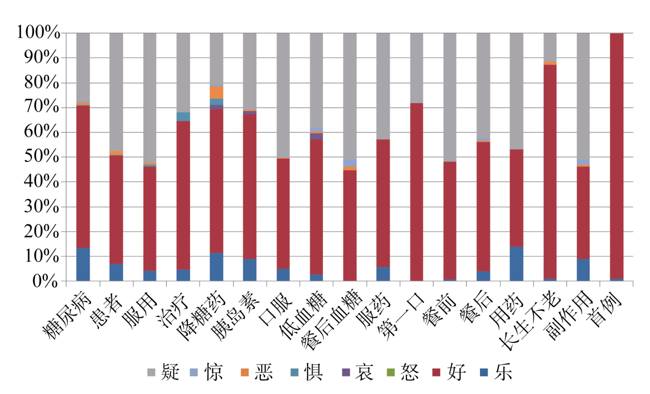


图 4 高频特征情感分布比例

4.3 数据结果验证

为了验证本文提出的情感分析方法的有效性，选取三名工作人员对数据进行人工标注，其中两人及以上标注结果相同的记录在案，标注结果不同的，三人商量之后决定结果。将自动分析的 7 类药物情感分类结果与人工标注的情感分类结果进行对比。

评价指标采用目前广泛接受的正确率(Precision)和召回率(Recall)，选用综合度量指标 F 值(F)作为 Precision 和 Recall 两者的调和平均数来衡量[34]评估分析的正确率。它们的计算公式如下所示。

$$\text{Precision}=\frac{\text{判断正确的该类别微博数}}{\text{判断为该类别的微博数}}$$

$$\text{Recall}=\frac{\text{判断正确的该类别微博数}}{\text{应判断为该类别的微博数}}$$

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$$

实验结果如表 11 所示。

表 11　实验性能评估分析

情感类别	Precision	Recall	F
乐	79.00%	83.15%	81.02%
好	77.18%	85.56%	81.15%
怒	85.73%	38.83%	53.45%
哀	83.05%	35.65%	49.89%
惧	53.42%	47.12%	50.07%
恶	64.67%	66.96%	65.80%
惊	54.58%	33.37%	41.42%
疑	77.33%	78.58%	77.95%

实验结果表明, 本文方法的正确率、召回率以及 F 值均较高, 其中"怒"和"哀"正确率最高, 而"乐"和"好"的召回率与 F 值均最高, 说明识别负面情绪的准确率较高, 而识别正面情绪的可靠度要高一些。本实验主要依赖于情感词典, 而情感词典的 8 类情感词的多少也是影响实验结果的关键因素, 所以应该尽可能地丰富情感词典, 减少对实验的影响。

5　结　语

微博的流行使得其中蕴含了丰富的情感信息, 通过对微博上的用户生成内容进行情感分析, 可以挖掘其中的商业价值和社会价值。目前, 情感分析领域的研究多是将情感进行正负二元分类, 或是加上中性三元分类, 很少更细致地划分情感类别, 也没有考虑情感的情感强度。本文针对人类情感复杂的特点, 在 DUTIR 情感词汇本体库的 7 类情感基础上丰富了一类占比较多的情感"疑", 通过语料分析及词典查阅构建了疑问词词表以此对 DUTIR 进行扩展, 让情感分类得更细腻, 同时考虑了表情符号对于情感表达的影响, 构建了表情符号情感词典, 参照表情符号在微博中的占比赋予其权值, 此外, 还构建了否定词词表与程度副词词表辅助情感分析, 有助于更准确地计算情感值, 从而得到每一类别情感的强度值, 便于对其进行比较分析。

此外, 针对微博的情感分析多应用于酒店、汽车等商务领域, 在药物方面的研究较少, 因此本文通过对 2 型糖尿病药物进行细粒度情感分析, 不仅得到了人们对于 7 类药物的情感分类及强度, 也得到了人们对于药物的哪种属性最为关心, 从而可为 2 型糖尿病的药物选择提供参考, 由本文结果可以得知二甲双胍类药物最为领先, 不仅所含情感丰富, 积极方面的情感也最多。其余各个药物均有自身的特点, 各有所长。

参考文献:

[1] CNNIC. 第 39 次中国互联网络发展状况统计报告[R]. 中国互联网络信息中心, 2017. (CNNIC.The Report of The 39th China Internet Development Statistics[R]. Information Center of the China Internet Network, 2017.)

[2] 蓝天广. 电子商务产品在线评论的细粒度情感强度分析[D]. 北京: 北京邮电大学, 2015. (Lan Tianguang. Fine-Grained Sentiment Analysis of E-Commerce Online Reviews [D]. Beijing: Beijing University of Posts and Telecommunications, 2015.)

[3] 李长江. 基于酒店中文评论情感倾向分析[D]. 广州: 华南理工大学, 2016. (Li Changjiang. Text Sentiment Polarity Analysis Based on Chinese Reviews in Hotel Domain [D]. Guangzhou: South China University of Technology, 2016.)

[4] 贾治中. 基于依存句法分析的中文评价对象抽取和情感倾向性分析[D]. 南京: 东南大学, 2016. (Jia Zhizhong.Chinese Opinion Target Extraction and Orientation Analysis Based on Syntactic Dependencies [D]. Nanjing: Southeast University, 2016.)

[5] 彭云, 万常选, 江腾蛟, 等. 基于语义约束 LDA 的商品特征和情感词提取[J]. 软件学报, 2017, 28(3): 676-693. (Peng Yun, Wan Changxuan, Jiang Tengjiao, et al. Extracting Product Aspects and User Opinions Based on Semantic Constrained LDA Model[J]. Journal of Software, 2017, 28(3): 676-693.)

[6] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment Classification Using Machine Learning Techniques[C]// Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Philadelphia. USA: Association for Computational Linguistics, 2002: 79-86.

[7] 杨艳霞. 基于分类的微博情感分析算法研究及实现[J]. 计算机与数字工程, 2017, 45(2): 197-200, 396. (Yang Yanxia. Microblog Sentiment Analysis Algorithm Research and Implementation Based on Classification [J]. Computer & Digital Engineering, 2017, 45(2): 197-200, 396.)

[8] 陈炳丰, 郝志峰, 蔡瑞初, 等. 面向汽车评论的细粒度情感分析方法研究[J]. 广东工业大学学报, 2017, 34(3): 8-14. (Chen Bingfeng, Hao Zhifeng, Cai Ruichu, et al. A

Fine-grained Sentiment Analysis Algorithm for Automotive Reviews [J]. Journal of Guangdong University of Technology, 2017, 34(3): 8-14.)

[9] 朱晓光. 基于半监督学习的微博情感分析方法研究[D]. 济南: 山东财经大学, 2014. (Zhu Xiaoguang. Research on Microblog Sentiment Analysis Based on Semi-supervised Learning [D]. Jinan: Shandong University of Finance and Economics, 2014.)

[10] 程佳军. 基于半监督递归自动编码的微博情感分析方法研究[D]. 长沙: 国防科学技术大学, 2014. (Cheng Jiajun. Research on Sentiment Analysis of Microblog Based on Semi-suprvise Recursive Auto Encoder [D]. Changsha: National University of Defense Technology, 2014.)

[11] 熊德兰, 程菊明, 田胜利. 基于HowNet的句子褒贬倾向性研究[J]. 计算机工程与应用, 2008, 44(22): 143-145. (Xiong Delan, Cheng Juming, Tian Shengli. Sentence Orientation Research Based on HowNet[J]. Computer Engineering and Applications, 2008, 44(22): 143-145.)

[12] 潘明慧, 牛耘. 基于多线索混合词典的微博情绪识别[J]. 计算机技术与发展, 2014, 24(9): 28-32, 36. (Pan Minghui, Niu Geng. Emotion Recognition of Micro-blogs Based on a Hybrid Lexicon[J]. Computer Technology and Development, 2014, 24(9): 28-32, 36.)

[13] 肖江, 丁星, 何荣杰. 基于领域情感词典的中文微博情感分析[J]. 电子设计工程, 2015, 23(12): 18-21. (Xiao Jiang, Ding Xing, He Rongjie. Analysis of Chinese Micro-blog Emotion Which Based on Field of Emotional Dictionary[J]. Electronic Design Engineering, 2015, 23(12): 18-21.)

[14] 王志涛, 於志文, 郭斌, 等. 基于词典和规则集的中文微博情感分析[J]. 计算机工程与应用, 2015, 51(8): 218-225. (Wang Zhitao, Yu Zhiwen, Guo Bin, et al. Sentiment Analysis of Chinese Micro Blog Based on Lexicon and Rule Set[J]. Computer Engineering and Applications, 2015, 51(8): 218-225.)

[15] 张珊, 于留宝, 胡长军. 基于表情图片与情感词的中文微博情感分析[J]. 计算机科学, 2012, 39(11A): 146-148, 176. (Zhang Shan, Yu Liubao, Hu Changjun. Sentiment Analysis of Chinese Micro-blogs Based on Emoticons and Emotional Words[J]. Computer Science, 2012, 39(11A): 146-148, 176.)

[16] 王文远, 王大玲, 冯时, 等. 一种面向情感分析的微博表情情感词典构建及应用[J]. 计算机与数字工程, 2012, 40(11): 6-9. (Wang Wenyuan, Wang Daling, Feng Shi, et al. An Approach of Building Microblog Smiley Emotion Lexicon and Its Application for Sentiment Analysis[J]. Computer & Digital Engineering, 2012, 40(11): 6-9.)

[17] 栗雨晴, 礼欣, 韩煦, 等. 基于双语词典的微博多类情感分析方法[J]. 电子学报, 2016, 44(9): 2068-2073. (Li Yuqing, Li Xin, Han Xu, et al. A Bilingual Lexicon-Based Multi-class Semantic Orientation Analysis for Microblogs[J]. Acta Electronica Sinica, 2016, 44(9): 2068-2073.)

[18] 何文娟. 微博情感营销对消费者购买意愿的影响研究[D]. 合肥: 安徽大学, 2016. (He Wenjuan. Research on the Influence of Microblog-Based Emotional Marketing on Consumers' Purchase Intention[D]. Hefei: Anhui University, 2016.)

[19] 史伟, 王洪伟, 何绍义. 基于微博情感分析的电影票房预测研究[J]. 华中师范大学学报: 自然科学版, 2015, 49(1): 66-72. (Shi Wei, Wang Hongwei, He Shaoyi. Study on Predicting Movie Box Office Based on Sentiment Analysis of Micro-blog[J]. Journal of HuaZhong Normal University: Natural Sciences, 2015, 49(1): 66-72.)

[20] 李鸣, 吴波, 宋阳, 等. 细粒度情感分析的酒店评论研究[J]. 传感器与微系统, 2016, 35(12): 41-43, 47. (Li Ming, Wu Bo, Song Yang, et al. Research on Hotel Reviews Based on Fine-grained Sentiment Analysis [J]. Transducer and Microsystem Technologies, 2016, 35(12): 41-43, 47.)

[21] 钱慎一, 杨铁松. 基于微博电影评论的情感分析研究[J]. 现代计算机(专业版), 2017(5): 48-51. (Qian Shenyi, Yang Tiesong. Research on Emotional Analysis Based on Micro-Blog Film Criticism [J]. Modern Computer, 2017(5): 48-51.)

[22] 赵晓航. 基于情感分析与主题分析的“后微博”时代突发事件政府信息公开研究——以新浪微博“天津爆炸”话题为例[J]. 图书情报工作, 2016, 60(20): 104-111. (Zhao Xiaohang. The Study on Government News Release in the Era of Post-microblog Based on Sentiment Analysis and Subject Analysis: A Case Study of the “Tianjin Explosion” on Sina Microblog[J]. Library and Information Service, 2016, 60(20): 104-111.)

[23] 缪茹一. 基于文本数据挖掘的微博情感分析与监控系统[D]. 杭州: 浙江工业大学, 2015. (Miu Ruyi. Microblog Sentiment Analysis and Monitoring System Based on Text Data Mining [D]. Hangzhou: Zhejiang University of Technology, 2015.)

[24] 崔安颀. 微博热点事件的公众情感分析研究[D]. 北京: 清华大学, 2013. (Cui Anqi.Study on Public Sentiment Analysis of Events in Microblogs [D]. Beijing: Tsinghua University, 2013.)

[25] 陈建美. 中文情感词汇本体的构建及其应用[D]. 大连; 大连理工大学, 2009. (Chen Jianmei. The Construction and

Application of Chinese Emotion Word Ontology[D]. Dalian: Dalian University of Technology, 2009.)

[26] 高宁. 现代汉语程度副词与否定副词共现的认知研究[D]. 长春: 吉林大学, 2013. (Gao Ning. A Cognitive Study on the Combination of the Degree Adverb and the Negative Adverb in Mandarin Chinese [D]. Changchun: Jilin University, 2013.)

[27] 施寒潇. 细粒度情感分析研究[D]. 苏州: 苏州大学, 2013. (Shi Hanxiao.Research on Fine-grained Sentiment Analysis [D]. Suzhou: Soochow University, 2013.)

[28] 陈国兰. 基于情感词典与语义规则的微博情感分析[J]. 情报探索, 2016(2): 1-6. (Chen Guolan. Microblog Sentiment Analysis Basing on Emotion Dictionary and Semantic Rule[J]. Information Research, 2016(2): 1-6.)

[29] 李婷婷, 姬东鸿. 基于SVM和CRF多特征组合的微博情感分析[J]. 计算机应用研究, 2015, 32(4): 978-981. (Li Tingting, Ji Donghong. Sentiment Analysis of Micro-blog Based on SVM and CRF Using Various Combinations of Features[J]. Application Research of Computers, 2015, 32(4): 978-981.)

[30] 马秉楠, 黄永峰, 邓北星. 基于表情符的社交网络情绪词典构造[J]. 计算机工程与设计, 2016, 37(5): 1129-1133. (Ma Bingnan, Huang Yongfeng, Deng Beixing. Generating Sentiment Lexicon of Online Social Network Based on Emotions[J]. Computer Engineering and Design, 2016, 37(5): 1129-1133.)

[31] 崔连超.互联网评论文本情感分析研究[D]. 济南: 山东大学, 2015. (Cui Lianchao. Research on Internet Review Text Sentiment Analysis [D]. Ji'nan: Shandong University, 2015.)

[32] 郑诚, 杨希, 张吉赓. 结合情感词典与规则的微博情感极性分类方法[J]. 电脑知识与技术, 2014, 10(13): 3111-3113, 3123. (Zheng Cheng, Yang Xi, Zhang Jigeng. Combining Emotional Dictionary and Rules of Microblogging Emotional Polarity Classification Method [J]. Computer Knowledge and Technology, 2014, 10(13): 3111-3113, 3123.)

[33] 汪会琴, 胡如英, 武海滨, 等. 2 型糖尿病报告发病率研究进展[J]. 浙江预防医学, 2016, 28(1): 37-39, 57. (Wang Huiqin, Hu Ruying, Wu Haibin, et al. Research Progress on Incidence of Type 2 Diabetes Mellitus[J]. Zhejiang Preventive Medicine, 2016, 28(1): 37-39, 57.)

[34] Li G, Hoi S C H, Chang K, et al. Microblogging Sentiment Detection by Collaborative Online Learning[C]//Proceedings of the 2010 IEEE International Conference on Data Mining, Sydney, Australia. USA: IEEE, 2010: 893-898.

作者贡献声明:

敦欣卉: 设计方案, 进行实验, 起草并修改论文;
张云秋: 确定论文选题, 完善研究方案, 论文最终版本修订;
杨铠西: 数据预处理, 分析功能的编程实现。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: 568977858@qq.com。

[1] 敦欣卉. 7 类药物情感分析结果.zip. 利用本文方法对 7 类药物进行情感分析的结果.

[2] 敦欣卉. 7 类药物人工分类结果.zip. 3 名工作人员对 7 类药物情感进行标注的最终结果.

[3] 敦欣卉. 7 类药物数据.zip. 从微博上爬取的含有 2 型糖尿病 7 类常用药物名称的微博内容.

[4] 敦欣卉. 表情符号词典.xlsx. 所构建的包含 39 个表情符号的表情词典.

[5] 敦欣卉. 表情符号词典 PMI 结果.xlsx. 39 个表情符号的 PMI 值列表.

[6] 敦欣卉. 表情符号词典构建数据.zip. 从微博上所爬取的含有 39 个表情符号的短文本.

[7] 敦欣卉. 高频特征词情感分析结果.zip. 利用本文方法对 17 个高频特征词进行情感分析的结果.

[8] 敦欣卉. 高频词微博数据.zip. 含有 17 个高频词的微博文本.

[9] 敦欣卉. 情感词汇本体.xlsx. 大连理工大学情感词汇本体库.

收稿日期: 2017-05-31
收修改稿日期: 2017-07-11

Fine-grained Sentiment Analysis Based on Weibo

Dun Xinhui[1] Zhang Yunqiu[1] Yang Kaixi[2]
[1](School of Public Health, Jilin University, Changchun 130021, China)
[2](International School of Information Science & Engineering, Dalian University of Technology, Dalian 116620, China)

Abstract: [**Objective**] This paper conducts a fine-grained sentiment analysis of Weibo posts by dividing the sentiments into eight categories and calculating their intensity values. [**Methods**] First, we analyzed the Weibo corpus to construct the question word list. Besides the seven sentiments defined by DUTIR, we added “suspected” to the list. Then, we used the Pointwise Mutual Information method, the impacts of negative words and the degree adverbs to construct the expression symbol dictionary. We employed Python to retrieve the needed data from Weibo, and applied the jiebaR package to segment the words. Finally, we classified the sentiments and calculated their intensity. [**Results**] We got the proportion of eight sentiment categories and sentiment intensity of commonly used drugs for diabetes. The Precision values of “angry” and “sad” were the highest (85.73% and 83.05%), while the Recall and F values of “happy” and “like” were the highest (more than 81%). The Precision, Recall and F values of “suspected” were 77.33%, 78.58% and 77.95% respectively. [**Limitations**] The sentiment dictionary needs to be expanded. [**Conclusions**] The proposed model could analyze the sentiment of Weibo Posts more effectively than traditional methods.

Keywords: Microblog Fine-grained Sentiment Analysis Drug

BBC Monitoring 加入 OpenAthens 联盟以扩展全球访问

近日, BBC Monitoring 加入了 OpenAthens 联盟, 进一步扩展其大学和其他国际组织的单点登录选择, 这些大学和国际组织将以全球 150 多个国家的多种语言访问 BBC Monitoring 的广播、新闻和媒体资源。

加入 OpenAthens 联盟将使 BBC Monitoring 能够在全球拓展其商业客户群。其商业客户目前包括: 媒体组织、外国政府、非政府组织、大学、大使馆、新闻机构、智库等。

“向英国和其他国家的高校提供服务是我们的商业战略。OpenAthens 意味着我们可以为那些希望为用户批量启用单点登录的客户提供简单又经过身份验证的访问。添加 OpenAthens 为英国以外的大学访问 BBC Monitoring 的订阅服务提供了另一个选择。”BBC Monitoring 业务发展总监 Markus Ickstadt 表示。

BBC Monitoring 为 BBC(英国广播公司)在英国和全球的新闻和节目团队、英国政府以及众多商业客户提供服务; 是 BBC 世界服务集团的一部分, 成立于 1939 年, 目前可以通过多种语言跟踪 150 多个国家, 并利用本地渠道来过滤、翻译和报告突发新闻、媒体行为和新兴趋势。

Eduserv 是一家位于英国巴斯的非营利 IT 服务公司, OpenAthens 是其一部分, 提供涵盖身份和访问管理、托管云服务、网络弹性和应用程序集成的一系列服务。OpenAthens 提供身份和访问管理解决方案, 并为 OpenAthens 联盟的成员提供支持服务。来自 46 个国家的 400 多个组织的全球 400 多万用户受益于 OpenAthens。

(编译自: http://www.bbc.co.uk/mediacentre/latestnews/2017/bbc-monitoring-openathens)

(本刊讯)